DOCUMENT RESUME

ED 377 201 TM 022 340

AUTHOR	Winfield, Linda
TITLE	Assessment, Equity, and Diversity in Reforming America's Schools. Project 2.5, Analytical Models for Performance and Delivery Standards.
INSTITUTION	Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	Nov 93
CONTRACT	R117G10027
NOTE	34p.; A version of this paper was presented at the CRESST conference "What Works in Performance Assessment?" (University of California, Los Angeles, CA, September 10, 1992).
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Cultural Differences; *Educational Assessment; Educational Change; Educational Improvement; Elementary Secondary Education; *Equal Education; Ethnic Groups; Minority Groups; *National Competency Tests; Outcomes of Education; Public Schools; Racial Differences; *Standards; Test Content; *Test Use; Urban Schools
IDENTIFIERS	*Reform Efforts

ABSTRACT

National standards and assessments are being proposed as a strategy for improving schools in the United States. However, proposed federal policies for implementation raise serious concerns about the extent to which national standards and assessments alone will help improve the quality of public education for all, or whether they will serve to deepen the already severe educational and economic cleavages that exist in this nation, especially along racial/ethnic lines. The implications of this policy for equity and diversity are examined in terms of antecedent instructional conditions, the proposed test, the testing context, and the diversity of learners to be assessed. Without a strong and serious commitment to opportunity to learn, this policy serves a symbolic and political function rather than an instrumental one in improving schooling outcomes, particularly for disadvantaged urban and racial/ethnic minority students. (Contains 93 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 377 201

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

National Center for Research on
Evaluation, Standards, and Student Testing

Final Deliverable – November 1993

Project 2.5 Analytical Models for Performance and Delivery Standards

Assessment, Equity, and Diversity in Reforming America's Schools

Linda Winfield, Project Director

U.S. Department of Education
Office of Educational Research and Improvement
Grant No. R117G10027 CFDA Catalog No. 84.117G

National Center for Research on
Evaluation, Standards, and Student Testing
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

TM022340

BEST COPY AVAILABLE

The work reported herein was supported under the Educational Research and Development Center Program cooperative agreement R117G10027 and CFDA catalog number 84.117G as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

The findings and opinions expressed in this report do not reflect the position or policies of the Office of Educational Research and Improvement or the U.S. Department of Education.

ASSESSMENT, EQUITY, AND DIVERSITY IN REFORMING AMERICA'S SCHOOLS[1]

Linda F. Winfield, CRESST/UCLA

Michael D. Woodard, UCLA Center for Afro-American Studies and Los Angeles Institute for MultiCultural Training

Abstract

National standards and assessments are being proposed as a strategy for improving schools in the United States. However, proposed federal policies for implementation raise serious concerns about the extent to which national standards and assessments alone will help improve the quality of public education for all, or whether they will serve to deepen the already severe educational and economic cleavages that exist in this nation, especially along racial/ethnic lines. We examine the implications of this policy for equity and diversity in terms of antecedent instructional conditions, the proposed test, the testing context, and the diversity of learners to be assessed. Without a strong and serious commitment to opportunity to learn, this policy serves a symbolic and political function rather than an instrumental one in improving schooling outcomes, particularly for disadvantaged urban and racial/ethnic minority students.

Since the 1930s, testing and assessment in America's schools have increased dramatically in response to demands for educational reform and accountability, whereas attention to curriculum has remained stable over that same period (Congress of the United States, 1992; Haney, Madaus, & Lyons, 1993). Most persons would agree that setting high educational standards and measuring students' performance against those standards are important processes. However, recent proposals for national standards and tests have been accompanied by considerable tension between the goals of quality and equality of opportunity. The national testing bill, Goals 2000: Educate America

[1] A version of this paper was presented at the CRESST conference "What Works in Performance Assessment?" September 10, 1992, University of California, Los Angeles. A version of the paper also appears in *Educational Policy*, *8*(1), March 1994.

Act (1993), is a clear example. The Goals 2000 bill relies on a top-down accountability model focused on testing new, world class standards and concentrates authority at the state level. Unfortunately, community- and school-based initiatives are ignored as a means for improving education, and equity and diversity are omitted from consideration in the Goals 2000 bill altogether.

This paper provides a framework by which to review equity, diversity, and assessment as essential elements of quality education and equality of opportunity. By equity, we mean the more or less similar distribution of financial and all other resources across schools so that each student can obtain an education required for meaningful participation in an increasingly technological society. By diversity or valuing diversity, we refer to creating an environment at the local school level in which every student, regardless of race/ethnicity, gender, ability, economic status, or national origin, has the opportunity to learn and achieve to his or her potential.

In this paper, we first briefly review components of the current national bill. Next, we examine equity issues in three areas: antecedent instructional conditions, the actual assessment, and the context for assessment. Because of the relevance of antecedent conditions to the issue of opportunity to learn, we examine this topic at some length. Finally, we examine diversity in terms of the characteristics of learners to be assessed.

The Problem: Reform Without Equity and Diversity

President Clinton's Goals 2000 bill is now offered as the newest strategy to improve schooling. As a strategy for reforming America's public schools, this bill concerns us in its failure to address two important concepts related to equality of opportunity in the United States—equity and diversity (Winfield & Woodard, 1992). In omitting a consideration of equity and diversity, will this bill help improve the quality of public education for all, or will this policy serve to deepen the already severe educational and economic cleavages that exist in this nation? There is reason for alarm. While high school completion rates among ethnic minority groups are increasing, college attendance among these groups is declining (Blackwell, 1991). Moreover, the unemployment rates among African American and Latino youth are double that of their White counterparts regardless of educational level (U.S. Department of Labor, 1984).

To the extent that an underlying objective of the national testing bill is to improve productivity and America's competitiveness in a global economy, then strategies must be targeted not only to middle-class America but also to those disenfranchised ethnic groups historically locked out of the American Dream.

The equity debate in education is not a new one, nor is diversity an issue novel to business and industry. The demand for racial equality received considerable attention in the educational reforms of the 1950s and during the Civil Rights Movement in the 1960s. For example, in *Brown v. Board of Education* (1954), the Kansas Board of Education decision led to the removal of legally sanctioned segregation and increased access to education for African Americans. Equal opportunity measures and affirmative action help address equity issues in employment. Diversity initiatives in education have received far less emphasis, however. Indeed, attention to equity and diversity issues was short-lived and has been reduced considerably since the 1970s (Orfield & Reardon, 1992; Wolf & Reardon, 1993). At the same ti:e, the continuing shift in the economy from manufacturing to more high-tech service jobs and the rapid changes in demographic characteristics of the workforce (Rumberger & Levin, 1987) make it imperative that a national education policy address forthrightly the issues of equity and diversity. By omitting these concepts from the proposal, the national testing bill provides empty promises for improving the quality of schooling and education. Any successful reform must include careful attention to persistent and systemic differences by race/ethnicity, gender, ability, or economic status in the distribution of opportunities, conditions, practices, and outcomes in schools and industry.

Proposed Use of Assessment

The proposed national testing bill (Goals 2000: Educate America Act, 1993) indicates that standards and testing will be used as the chief means to assess improvement in student learning at the local level. This bill is based in large part on the 1992 National Council on Education Standards and Testing (NCEST) report. Goals 2000 appears to be voluntary since it allows states to choose whether to submit their local exams for certification. But on the other hand, the bill makes it clear that it will be virtually impossible for states to participate in the bill's school improvement initiative or other federal initiatives without obtaining national testing certification. In theory, allowing

states to design test assessments locally that can be calibrated to national standards holds considerable promise; unfortunately, however, the technology and psychometrics for accomplishing this task do not exist currently, and years will be required for their development (see Baker & O'Neil, in press, for consideration of technical issues, design, analysis, and interpretation of performance assessments). Similar to reform measures of the past that focused on testing student outcomes, it is much easier for policy makers to use goals, standards, and tests as visible symbols of reform than to actually change the system of educational inequities that has existed over time (Giroux, 1992). The focus is on outcomes with little attention paid to the inputs and processes of schooling.

A few decades ago, a similar strategy was implemented as states and school districts used minimum competency tests in attempts to improve student outcomes. Political pressure and calls for accountability drove implementation despite the lack of evidence that these expensive testing reforms would improve learning (Winfield, 1990). Two important concepts were derived from this era, instructional and curricular validity. Instructional validity is defined as an actual measure of whether a school provides students with instruction in the knowledge and skills measured by a test (McClung, 1978). Simply put, was instruction provided to students that would enable them to perform successfully on a test? Curricular validity is the degree to which test items represent the objectives of the curriculum. A measure of curricular validity would be based on the degree to which a school uses available resources and appropriate methods necessary to teach objectives to specific student populations (Venezky, 1983). These two concepts originated from an important federal court ruling in Florida (*Debra P. v. Turlington*, 1979) that minimum competency tests (MCT) must be fair and must measure what had been taught. We question at this point whether schools could meet this criterion with the proposed higher standards and performance assessments.

Moreover, there are dramatic differences between minimum competency tests and the newer standards and assessments being proposed. First, MCTs measured minimum standards while the proposed new tests will measure more complex, higher order skills that are to be based on "world standards." New tests may not be paper-and-pencil exams but will require student

demonstrations and constructed responses and, therefore, rely on substantial interpretation by teachers (Herman, Aschbacher, & Winters, 1992). Because of the complexity of the tasks and teacher interpretation, demonstrating the instructional and curricular validity of newer assessments will be even more difficult than showing that students had opportunities to learn minimum standards.

The effectiveness of a test-driven strategy for improving schools is questionable because instructional conditions and practices must also be changed. As Madaus (1993) aptly states, "We cannot assess our way out of our educational problems" (p. 26). To illustrate the adverse impact of the national testing bill on students from non-European racial/ethnic groups, we next consider the equity issue in terms of the antecedent instructional conditions, the nature of the proposed test, and the context of testing.

Equity

Antecedent Instructional Conditions

The basic premise of the Goals 2000 proposal is that a system of national testing will reform schools and improve outcomes. By testing for important outcomes, schools and teachers will be held accountable, will change or adjust what they are doing in the classroom, and student motivation and learning will improve. This logic is fallacious on at least two counts. First, it ignores the gross inequities in antecedent instructional conditions that affect the learning of students from non-European racial/ethnic groups. Antecedent conditions include factors such as classroom and supplemental instruction, high school curriculum track, quality of teaching and counseling, and availability of social support services. These conditions usually mirror the caste-like status of these groups in American society and reflect not only financial inequities between school districts but also the opportunity to learn appropriate content, skills, and knowledge embodied in a test (Winfield, 1987, 1993).

Gross disparities in instructional conditions between racial/ethnic groups have been well documented. For example, one study found that the lack of counseling at high school entry is concentrated on students who are least likely to be able to use their families as an alternative source of information

(Lee & Ekstrom, 1987). In addition, disadvantaged, rural, and minority students are less likely to receive program planning counseling than their more advantaged and White counterparts (Lee & Ekstrom, 1987). Non-White students are disproportionately represented in lower nonacademic tracks, remedial classes, and special education classes where opportunity to learn is severely restricted (Braddock, 1990). Nationally, less adequate instructional materials are more likely found in schools where students are poor than in schools where students are wealthier (Barton, Coley, & Goertz, 1991; Kozol, 1991). Similarly, Oakes (1990) observed that students in minority schools have restricted access to "gatekeeping" courses such as algebra in junior high and calculus in senior high. According to Oakes (1990), teachers at these schools placed less emphasis on developing critical thinking and problem solving and offered fewer opportunities for students to become actively engaged in learning. Without adequate attention to these antecedent conditions that affect opportunity to learn, the proposed national exam unfairly penalizes students of color in financially strapped urban districts and results in "blaming the victim."

There are other problematic aspects of the Goals 2000 bill. For instance, the use of tests to change teaching and learning reflects an overreliance on top-down policy, an increasing distrust of professional judgment, and an attempt to re-assert political control of the schools. Elmore and McLaughlin (1988) note that educational reform operates on three levels, policy, administration, and practice, each with its own rewards, incentives, and limitations. They stated:

> Policy can set the conditions for effective administration and practice, but it can't predetermine how those decisions will be made. Administrative decisions can reflect policy more or less accurately and can set the conditions for effective practice, but it can't control how teachers will act in the classroom at a given point. Practice can reflect knowledge of more effective performance but this knowledge isn't always consistent with policy and administrative decisions. (p. v)

The three levels are loosely related. The authors argue that education reform must be grounded in an understanding of how teachers learn to teach, how school organizations affect practice, and how these factors affect children's performance. Practice is particularly important when considering the effect of testing on students from ethnic minority groups (Darling-

Hammond, 1993). Merely setting high standards and developing a new assessment system will not ensure changes in teacher behavior or student performance unless professional development activities and capacity building at the school level are given equal priority. From past experience, we know that if assessments are used in high-stakes situations, for example, for graduation or employment purposes, this will exacerbate the problem of student motivation and high school dropouts (Catterall, 1989; Kreitzer, Madaus, & Haney, 1989). The impact across states, districts, and schools will be determined by the level and quality of professional development.

Over a two-year period Aschbacher (1992) studied six sites attempting to implement alternative assessments. Two of the sites were large urban school districts interested in developing social studies assessment; the other four sites were an individual inner-city classroom, two small, districtwide reform efforts to assess performance in math, and a schoolwide reform effort. Teachers were provided with training in the rationale for alternative assessment, theories of learning and instruction that underlie the new approach, alternative assessment models and materials, and a process for developing performance assessments. Technical assistance consisted of several workshops totaling about 30 hours during a one-year period (one 3-day summer institute and two 1-day follow-up workshops). Data were collected in these sites using observations, interviews, and surveys. The populations of the schools varied in terms of socioeconomic and racial composition of students. Teachers used journals, open-ended questions, essays, and portfolios in either social studies or math.

Across the various sites, Aschbacher (1992) found similar barriers that hindered implementation of alternative assessments, but she also found similar factors that facilitated implementation. The major barriers observed included: teachers' use of the assessments primarily as learning activities rather than as a means to assess student performance; teachers' difficulty in specifying criteria for judging student work; teachers' fear and anxiety over assessment; and teachers' lack of time to learn, plan, practice, use, and reflect. Other barriers included the need for training and ongoing support and a lack of a long-range implementation plan.

The factors that appeared to facilitate the implementation of new assessments were: purposeful commitment by teachers to innovative

assessment and instruction; a provision for receiving training and technical assistance in a group; administrative support; and sustained technical assistance. Aschbacher noted that "working on alternative assessment led teachers to reflect more on their teaching practices, to consider the alignment of instruction and assessment, to view assessment as something positive that offers insights into how students think, and to see the importance of assessing growth and development" (p. 27). At the same time, Aschbacher cautions that it will take tremendous investments in time and externally provided professional development to implement quite modest alternative assessments. In addition, she states, "the kind of instruction that should support performance assessments is sorely lacking. We have observed great reluctance on the part of teachers to articulate desired student outcomes and to embrace the development of criteria and standards for assessment. Successful development and use of alternative assessments by teachers, therefore, requires a significant paradigm shift that cannot be sustained with just a few in-service meetings" (p. 27).

The world class standards suggested in Goals 2000 would allow schools and districts to gauge current practices, but there is a need to move beyond the test to professional development targeted towards intervention. Malcolm (1991) states:

> Does better assessment increase our responsibility for intervention, as better technology in medicine has increased the demand and the ethical dilemmas we face in determining the use of that technology in treatment? If we are prepared to do more, once we know more, perhaps the dangers of inequity possible in new assessment are worth the risk. But absent the resolve to intervene, one could argue that assessment becomes little more than voyeurism. (p. 31)

In order for an assessment to have a positive effect on learning and achievement of students from racial/ethnic minority groups, fundamental changes must simultaneously occur in the social organization of schools, school culture, teaching practices, and policies that have an effect on students' instructional conditions.

Measuring Opportunity to Learn

In the Goals 2000 bill, the heart of the debate on standards is the issue of equating standards with opportunity to learn. A potentially positive aspect of

standards as a proxy for opportunity to learn is that if such standards are fully implemented, then, for the first time in history, the inequities between schools and districts within states will have to be addressed (see Stevens, 1993a, 1993b for a discussion of these issues). Demonstrating that schools have met opportunity-to-learn standards will be no easy task, however. We will review methods for measuring the variable *opportunity to learn*, and then discuss studies of change and implementation that provide insight on this important topic in urban schools.

Most methods of measuring opportunity to learn are based on a coverage model; that is, how much of the curriculum has been covered or taught to students (Freeman, Belli, Porter, Floden, Schmidt, & Schwille, 1983; Freeman, Kuhs, Porter, Floden, Schmidt, & Schwille, 1983). There are several methods for obtaining estimates of the amount of content covered by a student or a group of students. Each method has advantages and limitations, and the purpose for assessing coverage should dictate the most appropriate method.

Direct observation and curriculum content analysis. Direct observation of classroom instruction by trained observers might yield the most valid measures of content covered. In one such naturalistic study, Barr (1973-1974) observed nine first-grade classrooms. She obtained a measure of the number of new words introduced in a specific time frame and the number of words learned by individual students. Although the direct observation method may provide the most valid measures of content covered, it is also time consuming and costly and may be best suited for well-funded research studies.

Another method of measuring content covered is to analyze the content of all curriculum materials used. Data could be obtained for the student or a particular group of students concerning initial and final placement in the curriculum, for example, number of pages covered. This method, however, does not typically include topics covered (or not covered) in class instruction or in textbooks and may not provide valid estimates of what students are actually taught.

In general, teachers determine the content that is taught. Brophy (1982) suggests that these decisions are likely to be influenced by external factors such as school and district objectives for standardized achievement tests, teachers' knowledge and beliefs about the particular content, and response to

individual differences among students. Moreover, only a subset of the intended curriculum is likely to be taught depending on available time, teacher experience, and skills in curriculum planning. What students actually learn is a reduced and somewhat distorted version of the intended curriculum. In one study that investigated the coverage of curriculum materials, the number of textbook pages covered by different fourth-grade mathematics classes was significantly related to achievement gain (Good, Grouws, & Beckerman, 1978). Similarly, the number of basals that first-grade reading groups completed was related to student achievement gain (Anderson, Evertson, & Brophy, 1979).

Teacher self-report. A third measure of content covered is teachers' self-reports. For example, in the CRAFT project, teachers were asked to keep detailed logs of time spent on reading and supportive activities (Harris & Serwer, 1966). Reading time (time spent directly teaching reading) was positively correlated with student achievement, although supportive time (time spent on discussion, writing, or audiovisual activities) and total time were not. Another method that relies on teachers' self-reports requires teachers to recall whether they have covered some specific content with a student or a group of students. A limitation of this method is the questionable accuracy of teachers' recollections of specific content covered with students over an entire school year. Results obtained from this method are less valid than are those from direct observation but are somewhat more accurate than those obtained through a content analysis of curriculum materials. Obtaining an estimate of content covered using teacher recall, however, may be more feasible for states and local school districts with limited resources.

Leinhardt and her colleagues (Cooley & Leinhardt, 1980; Leinhardt, 1983; Leinhardt & Seewald, 1981; Leinhardt, Zigmond, & Cooley, 1981) used the teacher self-report method extensively to obtain a measure of the degree to which test material had been taught. She suggests that teachers' self-reported estimates are reliable indicators of content covered. Teachers with 3 to 4 years of teaching experience possess accurate information about materials used in texts and about their own instructional practices (Leinhardt, 1983).

Cooley and Leinhardt (1980) asked teachers to estimate the percentage of students who had been taught the minimum material necessary to pass each item on a standardized achievement test in first- and third-grade reading and math. The pretest and teacher-reported estimates of test content covered

explained statistically significant portions of the variance in posttest achievement. In another study, teachers were asked to identify whether each student or a sample of students had been taught the information required to answer a test item. Information was also obtained on the degree of younger students' familiarity with the test's format. Data were collected on a student-by-item basis. The overlap between material taught and material tested was found to be a significant predictor of end-of-year reading achievement (Leinhardt et al., 1981).

In both studies, curriculum-based estimates were also measured using a content analysis of materials and teacher self-report methods. A comparison of the two measures indicated that the estimates based on the content analysis were less reliable than the teachers' self-reported estimates; however, each measure was equally useful in predicting posttest achievement (Leinhardt, 1983; Leinhardt et al., 1981).

Another limitation of the teachers' self-report method was that teachers' expectations about student competency may bias their estimates. Leinhardt (1983) found that teachers' expectations correlated with teacher-reported overlap but not when pretest information was included in the regression equation. "This means that teacher overlap estimates made at the end of the year do not simply reflect teacher expectations at the beginning of the year" (Leinhardt, 1983, p. 167). Indeed, analyses of teacher protocols indicated that teachers used a consistent search strategy to arrive at estimates rather than merely relying on personal perceptions (Leinhardt, 1983). Other researchers have also found that teachers' reports of material taught correlate substantially with achievement (Anderson, 1975; Chang & Raths, 1971; Husén, 1967; Lewy, 1972).

In a study of instructional conditions among first-grade students enrolled in Chapter 1 programs, Winfield (1987) found that a direct and positive relationship existed between the amount of coverage of specific standardized test objectives taught by classroom and Chapter 1 teachers and students' performance on standardized test items in reading. That is, Chapter 1 students performed as well as students in the national reference groups on items that both Chapter 1 and regular teachers rated as high in emphasis and coverage. On those items that teachers rated low in coverage and emphasis, Chapter 1 students performed lower than the national reference group. A

similar pattern of results was found in a study of content covered in fourth-grade mathematics (Winfield, 1993).

Measuring Opportunity to Learn—Qualitative Factors

Content covered is only one facet of opportunity to learn. The findings of a national study of promising programs in disadvantaged urban and rural schools suggest that opportunity to learn, defined as the actual curriculum students received, was influenced by factors such as level of implementation strategy, budgets, staff development, and administrative support (Stringfield, Millsap, Winfield, Brigham, Yoder, & Moss, 1992; Stringfield, Winfield, Millsap, Brigham, Gamse, & Moss, 1991). These sources of possible variations at the school and classroom levels make it important that opportunity-to-learn standards also include qualitative indicators of the school learning environment. Relying solely on quantitative indicators such as years of teaching experience, teacher certification, number of books in the library, and number of pages covered does not measure the actual use of resources and provides an incomplete picture of quality and opportunity to learn. More importantly, these indicators do not indicate the change processes that must occur in order for schools to improve instruction and learning in classrooms. Qualitative indicators that have been found to contribute to learning are more difficult to measure—factors such as interest and commitment of the teaching staff, the quality and impact of professional development activities, team building and collegiality, instructional leadership, and the existence of an academic culture conducive to learning on the part of teachers and students (Johnson, in press; Winfield, Johnson, & Manning, 1993). An assessment system that includes a clear and explicit design to measure the opportunity to learn will reveal inequities and ". . . inform policy and practice in teacher training, teaching practice, curricular design, and school organization" (Wolf & Reardon, 1993, p. 23).

In Chapter 1 elementary schoolwide project sites that experience small but steady gains in student achievement, changes in school and classroom conditions were systematically altered to improve the learning environment (Lytle, 1992; Winfield, 1991a; Winfield & Hawkins, 1993; Winfield, Hawkins, & Stringfield, 1992). These changes included shifting the locus of control from the district to principals and teachers in the school, changing professional

roles and responsibilities of teachers to include shared decision making and time for planning and reflection, changing the responsibilities of district personnel from supervising to intervention in the ongoing instructional program of the school, and changing the use of test scores and grades from collection to ongoing monitoring of student performance. These schools in extremely impoverished communities allocated resources to provide ongoing professional development and in-classroom support directly related to classroom instruction. Resources were allocated to implement incentives for teacher and student attendance and performance. Other conditions at the school level that had a positive effect on achievement included a provision for ongoing technical assistance, a working school leadership team, a system for monitoring and recognizing student progress, and a mechanism for involving parents. A shift occurred in school ethos and culture from a sense of hopelessness and failure to one of optimism and renewal. Over time, these schools experienced steady increases in attendance and achievement test outcomes. In this light, top-down policy suggested in Goals 2000 may not facilitate learning in poor urban areas. The major responsibility will fall at the state level, and few states have the capability or the commitment to assist districts in impoverished urban communities. Further, delivery standards and measures of opportunity to learn will be complex and difficult to obtain but are nonetheless necessary to ensure that all students have the opportunity to meet world class standards.

There is no consensus concerning what constitutes valid indicators of quality nor agreement on measurement (Burstein, 1993). At a minimum, we think the standards would include empirical data at the school level on financial resources available, staff and student and teacher assignment to classes, teacher turnover and teacher absenteeism, building level, district and state support for improvement, and the quality and support for professional development of teachers and principals. At the classroom level, information on curricular coverage and direct observations of teacher classroom practices are required. At the student level, information on classroom assignments and estimates of coverage of instructional materials are necessary. In short, ensuring the opportunity to learn requires a multifaceted quantitative and qualitative approach. National indicator systems being developed to validate curricula and learning opportunities include many of the components

addressed here (Burstein, Guiton, Mirocha, McDonnell, Ormseth, & Van Winkle, 1993; Porter, 1993).

The Actual Test

The new performance tests being proposed in Goals 2000 may be more appropriate for assessing learning compared to norm-referenced, standardized achievement tests because the new assessments would be direct measures of student learning. However, Wolf and Reardon (1993) cogently argue that performance assessments derive from a historical, philosophical, and political tradition in the U.S. that is antithetical to concerns for equity—a tradition that assumes intelligence is fixed and that excellence exists in only a few forms.

Common characteristics of performance measures are those that: (a) ask students to perform, create, or produce; (b) tap higher level thinking and problem-solving skills; (c) use tasks that represent meaningful instructional activities; (d) involve real-world applications; (e) rely on people and human judgment rather than machines to score; and (f) require new instructional and assessment roles for teachers (Herman et al., 1992).

A major concern is the accuracy with which a national test or system of examinations will measure the learning of students from non-European racial/ethnic groups (Winfield, 1992). Traditionally, the test validity question has been framed as a cultural bias issue, although several components related to the "testor" and "testee" have been documented (Johnson, 1987). Evidence for this notion has been difficult to substantiate. Miller-Jones (1989) suggests that the more appropriate argument regarding performance, culture, and testing lies not only in the *bias* features of the task but in the individual's interpretation of the task, which is related to previous cultural experiences. Performance differences are thus related to culturally determined ways of organizing information and solving problems (Miller-Jones 1989).

The assumption cannot be made, however, that alternative assessments would prevent unfairness or reduce achievement differences between racial/ethnic groups. The evidence collected thus far is inconclusive but suggestive of wider performance gaps between racial/ethnic groups. Badger (1993) found that the total test scores of students in low-SES schools were more than two standard deviations below those attained by students in high-SES

schools. However, students from low-SES as well as students from high-SES schools performed somewhat better on the open-ended questions than they did on the multiple-choice questions. Badger (1993) reports a similar pattern of responses for African American and Latino students and suggests that open-ended questions may give them a greater opportunity to respond. However, the relative performance gap remains between the racial/ethnic groups, and therefore the meaning of better performance on open-ended items is unclear. In contrast, the results of the 1992 National Assessment of Educational Progress (NAEP) in mathematics indicated that there was a larger gap between correct responses of White and Asian students and those of African American and Latino students for short constructed and extended response items as compared to the gap for multiple-choice items (Elliott, 1993). Moreover, an examination of performance on NAEP open-ended essay exams and multiple-choice reading tests shows that achievement differences between African American and White students are the same regardless of test type (Baker, O'Neil, & Linn, 1993; Linn, Baker, & Dunbar, 1991). The importance of type of test as an explanation of racial/ethnic group score differences is unclear. More importantly, alternative assessments by themselves are no panacea for needed changes in schooling.

In a study that examined the relationship between portfolio scores and standardized test performance, large discrepancies occurred in the identification of Chapter 1 students. The correlation between reading portfolios and standardized reading tests was .55 and between math portfolios and math standardized tests was .66 (Colwell & Mitchell, 1993). When teacher-judged portfolio scores were compared to standardized test scores, there was considerable discrepancy in classifying students. More students were perceived by their teachers as performing better than was indicated by standardized tests (Colwell & Mitchell, 1993). Similar discrepancies occur with respect to students from racial/ethnic groups. LeMahieu (cited in Madaus, 1993) found that African Americans received lower scores on their portfolio evaluations than Whites regardless of the race of the rater. On another long-term independent writing assignment, more than 70% of those classified as highly proficient writers were White while more than 80% of low-proficiency writers were Black. When the writing from the portfolios was compared with this independent record, the highly proficient writers on the

independent measure scored even higher on the portfolio samples. The difference appeared to be the self-selection of materials. African American students tended not to choose material from their portfolios that presented their best writing (LeMahieu as cited in Madaus, 1993). The question must be raised, however, as to whether "best writing" is to some extent culturally determined.

Despite the concerns, when used with other measures, performance tasks within the classroom can be useful for diagnosing and assessing individual student progress. In a national or even regional examination context (Resnick & Resnick, 1992), however, these measures pose problems of generalizability, validity, and subjective bias in judging performance of students from racial/ethnic and class groups. For example, on a math test for proposed for 10th-grade students, there is no correct or incorrect answer for a math item (Chira, 1991). Students are required to write a report and include recommendations concerning whether to buy or lease cars. As a worker in a corporation purchasing department, a student is provided information on financing terms and interest rates. First students must understand how to do the calculations in order to compare alternatives. Given correct calculations, if students were to select the more costly alternative because they value automobile ownership/leasing or spending rather than saving money, would such a response be acceptable? To what extent would the content of the item—corporations, purchasing department, and leasing—provide an advantage for middle-class, White, 10th-grade students? Developers of a national test must be prepared to incorporate a *multicultural orientation in the development, standards, and criteria*. An important issue is whose content gets included! If students are required to construct a response in written or oral form, what content is appropriate/acceptable?

Questions about the validity and appropriateness of a national test focus not only on the selection of content but also on the standard being applied. For example, in a study of performance-based literacy tasks taken from the NAEP Young Adult Literacy Assessment (Kirsch & Jungteblut, 1986), eighth-grade, inner-city, African American students were administered tasks in a one-to-one situation and asked to "think aloud" about how they would go about solving the tasks (Winfield, 1991b). One task included a poem that described a scenario for an individual named Joe and alluded to death, the metal barrel of a gun,

and other war paraphernalia. It might be obvious to an adult reader that the passage referred to someone preparing to go to war. When one youngster was asked to explain his interpretation, he replied, "He's getting ready to go out in the street." The student was asked to elaborate and replied, "He got a gun . . . people get killed in the street where I live." This youngster, growing up in a violent, inner-city neighborhood where innocent children are wounded and killed by stray bullets, had read, interpreted, and constructed a response based on his experience and background knowledge. When the youngster and test administrator re-examined the passage together and looked for other clues, the student was able to obtain the socially correct response. The point is that the student's initial response would have been judged unacceptable and incorrect.

Successful completion of performance-based tasks will be heavily influenced by culture and opportunity to learn specific content—most of which will reflect European cultures. Providing a detailed and rich context for the assessment may allow some students to demonstrate better performance than what might occur on a multiple-choice measure. However, unless the contexts are derived from a multicultural perspective, familiarity with the context will still favor certain racial/ethnic groups, providing them an advantage. Moreover, the subjective bias inherent in judging or rating students' oral or written performances will influence the validity of these measures. Thus, it is likely that on performance-based measures, the achievement gap between subgroups will remain or increase as some of the early studies have shown.

Historical Context

In the United States, testing increased in the early 20th century when attendance in school was made compulsory and educators needed ways to deal with the influx of immigrant students. Emphasis was placed on selecting individuals for available educational opportunities rather than maximizing students' potential success in such opportunities. For African Americans and Latinos, tests have been used primarily to perpetuate myths of inferiority and restrict access rather than to select educational opportunities. Many scholars have noted the ideological basis of IQ testing and how this notion was used to provide scientific legitimacy for the belief in racial group differences in

intelligence (Sewell, Ducette, & Shapiro, 1991). The gatekeeping function of tests in American society has been documented recently also (National Commission on Testing and Public Policy, 1990).

The historical context and legacy of testing in this country, combined with a lack of concern for equity, mitigate against any change in the context in which test results are used. A new and improved assessment will not automatically change pre-existing inequities in instructional and social conditions among underrepresented groups. Moreover, from recent history, not in the area of assessment but in the evaluation of performance of African American children, context interacted with the accuracy of the measure. Research on language performance conducted in laboratory settings in the late 1960s and early 1970s led to conclusions that African American children had "no language" and were "non verbal" (Osser, Wang, & Zaid, 1969; Weener, 1969). These measures of performance were conducted with children in unfamiliar laboratory settings with unfamiliar experimenters from different racial groups. Young children were asked to respond to verbal or written stimuli, repeat phrases or sentences, or answer structured questions. In many instances, the students responded in monosyllables. When this scenario occurred, researchers concluded that poor African American children were nonverbal and had no language. In many of these studies, the context interacted with characteristics of the learner to severely depress children's performance. Other researchers changed the context and demonstrated the ethnocentric bias in much of this research and the need to address ecological validity—that is, the need for studies of language performance in naturalistic settings, which provided a much richer view of the verbal performance and capabilities of children (Baratz & Baratz, 1970; Labov, 1972).

These contextual issues also suggest reason to worry about the reliability and accuracy of ratings, scores, and judgments made of the performance of students from various racial/ethnic groups. Even for students not from these groups, the reliability of portfolio scores in reading and math in one state program has been found to be quite low (Koretz, 1993). Performance assessed through a demonstration or exhibition is heavily influenced by students' verbal skills such that dialect or accent may influence raters' scores. Individuals who speak in dialects or with accents are more likely to be judged as less intelligent and less capable. Tucker (1979) found that teachers' judgments of

students who used Black English Vernacular were generally negative. These students were rated as being less intelligent, less competent, and less capable of succeeding in school.

The context from both a historical and situational perspective suggests caution in developing and implementing performance assessments, particularly when high stakes are attached to the decision. We suggest that a useful context for developing such measures is at the local rather than the national level, where these measures could be fully integrated with ongoing professional development of teachers and local curriculum development. In this approach, we would use no ordinary microscope, but one with double or perhaps triple lenses whereby principals, teachers, parents, and students can have a voice in the process. The use of such assessment information at the local school level has a greater probability for generating the kind of data necessary to change and understand achievement in schools. Moreover, individuals at the school and district levels informed by such data are in a better position to intervene to produce the types of changes needed to improve achievement. To ensure that high standards are implemented for all students, the emphasis would be placed not only on national world class standards or subject matter standards such as those developed by NCTM but also on the actual curriculum as informed by these groups and local constituencies. However, unless commensurate attention and funding is available for local capacity building and professional development, it is a poor investment to spend millions of dollars on developing national standards or assessments.

Diversity

Learner Characteristics

Changes in the structure of the economy and the demographics of the workforce provide a real opportunity to assess whether this nation can live up to its ideal of equality, a society where diversity is valued so that race/ethnicity and gender are not artificial barriers to educational achievement and economic success. America's economic base is enhanced when women, people of color, the differently abled, and older Americans can reach their full potential in education and the workforce. By the year 2000, the economy will

grow at a relatively healthy pace, and the workforce will grow slowly. Native Euro-American males will make up only 15% of the new labor market entrants compared to the 47% in that category today. In contrast to Euro-American males, people of color will double their share of the labor force to make up 29% of the new entrants (U.S. Department of Labor, 1988). At first glance, the greater share of a more slowly growing workforce suggests improvement for the employment prospects for workers of color. In major urban areas, however, the proposed national exams, if used in a high-stakes manner, will serve as an additional barrier to employment opportunities of African Americans and Latinos unless commensurate investments are made in education and training. In central city areas, the inputs into education have been less than in suburban areas, and semiskilled manufacturing jobs steadily give way to high-skilled service jobs (Wilson, 1987). The adverse effect of the skills mismatch on less educated and less skilled workers has been documented. This mismatch also contributes to the gap between Black-White earnings and income differentials (Mincy, 1991). The shrinking number of younger Euro-American males, the rapid pace of industrial change, and the ever-increasing skill requirements make the task of fully preparing and utilizing workers of color particularly urgent between now and the year 2000. Therefore, in addition to instructional conditions, the actual test, and the context of testing, a national examination system must also consider characteristics of the learners being assessed.

Immigrants. The 1990 Census showed that 9 million people emigrated to the U.S. during the 1980s, and by the year 2000, immigrants will represent the greatest share of the increase in the population and the workforce since World War II. Even with immigration law that now emphasizes access for the skilled and professionally trained, approximately 750,000 legal and illegal immigrants are projected to enter the United States annually for the remainder of the century. Two-thirds or more of immigrants of working age are likely to join the workforce (Johnston & Packer, 1987). The greatest impact of immigration will be felt in the South and West, ports of entry and areas of immigrant concentration. Indeed, the influx of immigrants is expected to drastically reshape local economies, promote faster economic growth, and create labor surpluses along with placing severe demands on schools.

Therefore, some regional adjustments must be made to facilitate instruction and assessment for citizenship and work.

Taken together, these demographic changes will mean that students in our public schools and the new workers entering the workforce by the year 2000 will be very different from those of today and yesterday. People of color, women, and immigrants will make up more than five-sixths of the net additions to the workforce between now and the year 2000, though they make up only about half of it today (Johnston & Packer, 1987). For the great majority of immigrants, English is not the primary language. The more than 2 million immigrant youth who enrolled in U.S. schools over the past decade represent a great challenge, not only because of limited English proficiency, but also because many have had little or no formal schooling in their native countries (McDonnell & Hill 1993). Efforts to reform education, including the development of national tests and standards, typically ignore the special needs of students with limited English proficiency (August & Hakuta 1993; McDonnell & Hill 1993). All students are implied to have the same educational needs and the same access to education (Alvarez & Hakuta, 1992). If a national examination system is implemented, one can easily imagine that a dual track educational structure will be forthcoming—one in which immigrant and ethnic/minority group students are disproportionately represented in the bottom.

Immigrant populations and non-English speakers are vulnerable to being unfairly evaluated by performance assessments. Linguistic and sociocultural background characteristics interact with test performance and influence teacher decisions and beliefs about students' capabilities (O'Connor, 1989). There are serious concerns about the questionable reliability and accuracy of tasks when one considers language proficiency and other cultural characteristics of the particular learners (Estrin, 1993). For example, when asked to perform verbal tasks, non-native English speakers might require additional time for processing the task. Additional time within a classroom setting does not present a problem; however, it will be an issue in assessing the comparability of performance scores across classrooms, schools or districts. Some of the past psycholinguistic research on bilingualism (Kolers, 1968; Lambert, 1972; McNamara, 1967) may inform this issue. Similarly, teachers and assessors need to understand and know how linguistic complexity of

verbal tasks influences difficulty levels and interacts with the range of capabilities of non-English speakers. This type of basic information is required in order to ensure fair and accurate performance assessments for non-English speakers.

Conclusion

Equity and valuing diversity are necessary components of any educational policy that ensures that each American, regardless of race/ethnicity, gender, or national origin, can obtain the education required to be productive in an increasingly technological society. The national standards and assessment proposed in Goals 2000 will not effectively change inequitable education and employment opportunities, in part because they focus primarily on the outcomes of schooling. In fact, unless commensurate effort goes into addressing antecedent instructional conditions and guarantees provided for opportunity-to-learn standards, the bill will actually exacerbate existing inequalities by creating additional barriers and limiting opportunities for upward mobility out of poverty. As a result, America's problems including inadequate productivity, unemployment, crime, and dependency will continue to increase. A greater investment in education and training at all levels is needed to assure that employers have a qualified workforce in the years after 2000 and to finally deliver on America's unfulfilled promise—equality of opportunity. As currently construed, national standards and assessment will only ensure that those students and individuals who have historically been disenfranchised and underrepresented remain in a subordinated position and bear the burden of proposed school reform.

The last three decades of testing have not led to dramatic improvements in the educational system, particularly for students in financially strapped urban districts. Newer types of assessments are promising as measures of how students learn; however, the use of such tests as a policy tool carries certain risks (Haertle, 1989). Changes in national standards and assessment are not the necessary conditions for improving student and school achievement. Policies and practices that directly address conditions of current inequities in opportunities to learn at the school, district, and state levels have a greater probability of improving school learning and achievement. Such policies include equitable school financing, funding curriculum development,

increasing training and staff development for teachers and administrators in content area assessments, and improving assessment course content and requirements in universities. These policies, which affect teaching and learning, are more closely related to practices in schools and classrooms. Additionally, investments in local research and development units to expand types of tests used, and collaborative ventures between schools and industry and between schools and research and development centers are viable alternatives to improve assessment practices and use in the nation's schools. Even these partial solutions are insufficient to guarantee that equity and diversity issues will be considered. Only when policy makers consider opportunity-to-learn standards as important as implementing national standards and assessment will we ensure that those students and individuals historically disenfranchised will share in the American dream of opportunity for educational achievement and economic success.

References

Alvarez, L. P., & Hakuta, K. (1993). Enriching our views of bilingualism and bilingual education. *Educational Researcher, 21*(2), 4-9.

Anderson, L. (1975). *Opportunity to learn, test bias and school effects.* Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Anderson, L., Evertson, C., & Brophy, J. (1979). An experimental study of effective teaching in first-grade reading groups. *Elementary School Journal, 79*, 192-223.

Aschbacher, P. R. (1992). *Issues in innovative assessment for classroom practice: Barriers and facilitators* (CSE Tech. Rep. No. 359). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

August, D., & Hakuta, K. (1993). *Federal education programs for the limited English proficiency students: A blue print for the second generation* (Report for the Stanford Working Group). Stanford, CA: Stanford University.

Badger, E. (1993, March). *The role of expectations and fairness in statewide assessment programs: Lessons from Massachusetts.* Paper presented at the Ford Foundation Symposium on Equity and Educational Testing and Assessment, Washington, DC.

Baker, E. L., & O'Neil, H. F., Jr. (in press). Performance assessment and equity: A view from the U.S.A. *Assessment in Education.*

Baker, E. L., O'Neil, H. F., Jr., & Linn, R. L. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist, 48*, 1210-1218.

Baratz, S. S., & Baratz, J. (1970). Early childhood intervention: The social science base of institutional racism. *Harvard Education Review, 40*, 29-50.

Barr, R. (1973-1974). Instructional pace differences and their effect on reading acquisition. *Reading Research Quarterly, 9*, 526-554.

Barton, P. E., Coley, R. J., & Goertz, M. E. (1991). *The state of inequality* (Policy Information Report). Princeton, NJ: Educational Testing Service, Policy Information Center.

Blackwell, J. (1991). Blacks and Hispanics in the educational pipeline. In G. E. Thomas (Ed.), *U.S. race relations in the 1940s and 1990s: Challenges and alternatives.* New York: Hemisphere.

Braddock, J. H. (1990). *Tracking: Implications for African American students.* Baltimore, MD: Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Brophy, J. (1982). How teachers influence what is taught and learned in classrooms. *Elementary School Journal, 83*(1), 1-13.

Brown v. Board of Education, 347 U.S. 483 (1954).

Burstein, L. (1993, April). *Validating national curriculum indicators: A conceptual overview of the RAND/CRESST NSF project.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Burstein, L., Guiton, G., Mirocha, J., McDonnell, L., Ormseth, T., & Van Winkle, J. (1993, April). *Validating national curriculum indicators: Student work samples as benchmarks of learning opportunities and instructional practices.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Catterall, J. S. (1989). Standards and school dropouts: A national study of tests required for graduation. *American Journal of Education, 98*(1), 1-34.

Chang, S. S., & Raths, J. (1971). The school's contribution to the cumulating deficit. *Journal of Educational Research, 64*, 272-276.

Chira, S. (1991, August 7). Educators draw outline for nationwide testing. *New York Times*, p. A19.

Colwell, C. D., & Mitchell, D. E. (1993). *A district implementation of performance-based and portfolio assessment for compensatory education students.* Riverside: University of California, California Educational Research Cooperative.

Congress of the United States. (1992). *Testing in America's schools: Asking the right questions* (OTA-SET 520). Washington, DC: Office of Technology Assessment.

Cooley, W. W., & Leinhardt, G. (1980). The instructional dimensions study. *Educational Evaluation and Policy Analysis, 2*(1), 7-25.

Darling-Hammond, L. (1993, March). *Equity issues in performance-based assessment.* Paper presented at the Ford Foundation Symposium on Equity and Educational Testing and Assessment, Washington, DC.

Debra P. v. Turlington, 474 F. Supp. 244, 252 (M.D. Florida 1979).

Elliott, E. J. (1993, March). *National testing and assessment strategies: Equity implications of leading proposals for national examinations.* Paper

presented at the Ford Foundation Symposium on Equity and Educational Testing and Assessment, Washington, DC.

Elmore, R., & McLaughlin, M. (1988). *Steady work, policy, practice, and the reform of American education* (R-3574-NIE/RC, p. v). Santa Monica: RAND.

Estrin, E. T. (1993). *Alternative assessment: Issues in language, culture and equity.* San Francisco, CA: Far West Laboratory.

Freeman, D. J., Belli, G. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). The influence of different styles of textbook use on instructional validity of standardized tests. *Journal of Educational Measurement, 20*(3), 259-270.

Freeman, D. J., Kuhs, T. M., Porter, A. C., Floden, R. E., Schmidt, W. H., & Schwille, J. R. (1983). Do textbooks and tests define a national curriculum in elementary school mathematics? *Elementary School Journal, 83*(5), 501-513.

Giroux, H. A. (1992). Educational leadership and the crisis of democratic government. *Educational Researcher, 21*(4), 4-11.

Goals 2000: Educate America Act, H.R. 92, House of Representatives, January 5, 1993.

Good, T., Grouws, D. A., & Beckerman, T. M. (1978). Curriculum pacing: Some empirical data in mathematics. *Journal of Curriculum Studies, 10,* 75-82.

Haertle, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 25-50). Boston: Kluwer Academic Publishers.

Haney, W. M., Madaus, G. F., & Lyons, R. (1993). *The fractured marketplace for standardized testing.* Boston: Kluwer.

Harris, A. J., & Serwer, B. L. (1966). The CRAFT project: Instructional time in reading. *Reading Research Quarterly, 2,* 27-56.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of twelve countries, Vol. 2.* New York: Wiley.

Johnson, R. (in press). Restructuring schools: The debate. In L. Rendon (Ed.), *Educating the new majority.* San Francisco: Jossey-Bass.

Johnson, S. (1987). Test fairness and bias: Measuring academic achievement among Black youth. *The Urban League Review*, *11*(1), 76-92.

Johnston, W. B., & Packer, A. H. (1987). *Workforce 2000*. Washington, DC: Hudson Institute.

Kirsch, I. S., & Jungteblut, A. (1986). *Literacy: Profiles of America's young adults*. Princeton, NJ: National Assessment of Educational Progress.

Kolers, P. A. (1968). Bilingualism and information processing. *Scientific American*, 78-85.

Koretz, D. (1993). *Interim report: The reliability of the Vermont portfolio scores in the 1992 school year* (CSE Tech. Rep. No. 370). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Kozol, J. (1991). *Savage inequalities*. New York: Crown.

Kreitzer, A., Madaus, G., & Haney, W. (1989). Competency testing and dropouts. In L. Weis, E. Farrar, & H. Petrie (Eds.), *Dropouts from school: Issues, dilemmas, and solutions* (pp. 129-152). Albany: State University of New York Press.

Labov, W. (1972). *Language in the inner city*. Philadelphia: University of Pennsylvania Press.

Lambert, W. E. (1972). Psychological studies of the interdependencies of the bilingual's two languages. In W. E. Lambert (Ed.), *Language, psychology and culture* (pp. 300-330). Stanford, CA: Stanford University Press.

Lee, V. E., & Ekstrom, R. B. (1987). Student access to guidance counseling in high school. *American Educational Research Journal*, *24*(2), 287-310.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In G. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 151-170). Boston: Kluwer-Nijhoff.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught. *Journal of Educational Measurement*, *18*(2), 85-96.

Leinhardt, G., Zigmond, N., & Cooley, W. (1981). Reading instruction and its effects. *American Educational Research Journal*, *18*(3), 343-361.

Lewy, W. (1972). Opportunity to learn and achievement in three subject matter areas. *Journal of Experimental Education*, *41*, 68-73.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991, Winter). Complex, performance-based assessment: Expectations and validation criteria. *Evaluation Comment*, pp. 2-9.

Lytle, J. H. (1992). Prospects for reforming urban schools. *Urban Education*, *27*(2), 109-131.

Madaus, G. (1993, March). *A technological and historical consideration of equity issues associated with proposals to change our nation's testing policy*. Paper presented at the Ford Foundation Symposium on Equity and Educational Testing and Assessment, Washington, DC.

Malcolm, S. (1991). Equity and excellence through authentic science assessment. In G. Kulm & S. Malcolm (Eds.), *Science assessment in the service of reform* (p. 316). Washington, DC: American Association for the Advancement of Science.

McClung, M. S. (1978). Competency testing programs: Legal and educational issues. *Fordham Law Review*, *47*, 651-712.

McDonnell L. M., & Hill P. T. (1993). *Newcomers in American schools: Meeting the needs of immigrant youth*. Santa Monica, CA: RAND.

McNamara, J. (1967). The bilingual's linguistic performance: A psychological overview. *Journal of Social Issues*, *23*, 58-77.

Miller-Jones, D. (1989). Culture and testing. *American Psychologist*, *44*(2), 360-366.

Mincy, R. B. (1991). Workforce 2000, silver bullet or dud?: Job structure changes and economic prospects for Black males in the 1990s. *Challenge*, *2*(1), 50-76. [Atlanta, GA: Morehouse Research Institute]

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America. Report of the National Commission on Testing and Public Policy*. Chestnut Hill, MA: Boston College.

National Council on Education Standards and Testing. (1992). *Raising standards for American education. A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people.* Washington, DC: U.S. Government Printing Office.

Oakes, J. (1990). *Multiplying inequalities: The unequal distribution of mathematics and science opportunities.* Santa Monica, CA: RAND.

O'Connor, M. C. (1989). Aspects of differential performance by minorities on standardized tests: Linguistic and sociocultural factors. In B. Gifford (Ed.), *Test policy and test performance: Education, language, and culture* (pp. 129-181). Boston: Kluwer.

Orfield, G., & Reardon, S. F. (1992, September). *Separate and unequal schools: Political change and the shrinking agenda of urban school reform*. Paper

prepared for the annual meeting of the American Political Science Association, Chicago, IL.

Osser, H., Wang, M. D., & Zaid, F. (1969). The young child's ability to imitate and comprehend speech: A comparison of two sub-cultural groups. *Child Development*, *40*, 1063-1076.

Porter, A. C. (1993, June). School delivery standards. *Educational Researcher*, pp. 24-30.

Resnick, L., & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. C. O'Connor (Eds.), *Future assessments: Changing views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer.

Rumberger, R. W., & Levin, H. M. (1987). *Forecasting the impact of new technologies on the future job market* (Project Report No. 84A4). Stanford, CA: Stanford University, Institute for Educational Finance and Governance.

Sewell, T. E., Ducette, J. P., & Shapiro, J. (1991, August). *Cultural diversity and educational assessment.* Paper presented at the annual meeting of the American Psychological Association, San Francisco.

Stevens, F. I. (1993a). *Defining and analyzing opportunity to learn in U.S. public schools: Issues of equity for poor and minority students.* Washington, DC: National Center for Education Statistics.

Stevens, F. I. (1993b). The opportunity to learn: Implications for school reform. *The Journal of Negro Education*, *62*(3), 227-248.

Stringfield, S., Millsap, M. A., Winfield L., Brigham N., Yoder, N., & Moss, M. (1992). *Urban and suburban/rural special strategies for educating disadvantaged children. Second year report* (Contract No. LC 90010001, LC 9001002). Baltimore, MD: The Johns Hopkins University.

Stringfield, S., Winfield, L., Millsap, M. A., Brigham, N., Gamse, B., & Moss, M. (1991). *Urban and suburban/rural special strategies for educating disadvantaged children. First year report* (Contract No. LC 90010001, LC 9001002). Baltimore, MD: The Johns Hopkins University.

Tucker, C. (1979). *Teachers' perception of Black English Dialect and variants.* Unpublished doctoral dissertation, University of North Carolina, Chapel Hill.

U.S. Department of Labor, Bureau of Labor Statistics. (1984). *Current population survey*. Washington, DC: U.S. Government Printing Office.

U.S. Department of Labor, Employment Standards Administration. (1988). *Opportunity 2000: Creative affirmative action strategies for a changing workforce, 1-18.* Washington, DC: U.S. Government Printing Office.

Venezky, R. L. (1983). Curricular validity: The case for structure and process. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 181-195). Boston: Kluwer-Nijhoff.

Weener, P. D. (1969). Social dialect differences and the recall of verbal messages. *Journal of Educational Psychology, 60,* 194-199.

Wilson, W. J. (1987). *The truly disadvantaged.* Chicago, IL: University of Chicago Press.

Winfield, L. F. (1987). Teacher's estimates of test content covered and first grade students' reading achievement. *Elementary School Journal, 4,* 437-454.

Winfield, L. F. (1990). School competency testing reforms: Exploring a national perspective. *Educational Evaluation and Policy Analysis, 12*(2), 157-173.

Winfield, L. F. (1991a). Lessons from the field: Case studies of evolving schoolwide projects. *Educational Evaluation and Policy Analysis, 13*(4), 353-362.

Winfield, L. F. (1991b, December). *Understanding adult and children's literacy proficiency.* Paper presented at the National Reading Conference, Palm Springs, CA

Winfield, L. F. (1992). *Cautionary notes in the development of performance based assessments: The impact on racial/ethnic minority groups.* Testimony prepared for hearings, House Education and Labor Committee.

Winfield, L. F. (1993). Investigating test content and curriculum content overlap to assess opportunity to learn. *The Journal of Negro Education, 62*(3), 288-323.

Winfield, L. F., & Hawkins, R. (1993). *Longitudinal effects of Chapter 1 schoolwide projects on disadvantaged students* (CDS Rep. No. 46). Baltimore, MD: The Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Winfield, L. F., Hawkins, R., & Stringfield, S. (1992). *A description of Chapter 1 schoolwide projects and effects on student achievement in six case study schools* (CDS Rep. No. 37). Baltimore, MD: The Johns Hopkins University, Center for Research on Effective Schooling for Disadvantaged Students.

Winfield, L. F., Johnson, R., & Manning, J. (1993). Managing instructional diversity. In P. Forsyth & M. Tallerico (Eds.), *City schools: Leading the way* (pp. 97-130). Newbury Park, CA: Corwin Press.

Winfield, L. F., & Woodard, M. D. (1992, January 31). Where are equity and diversity in Bush's America 2000? *Education Week*, pp. 31, 33.

Wolf, D. P., & Reardon, S. F. (1993, March). *Equity in the design of performance assessments: A handle to wind up the tongue with?* Paper prepared for the Ford Foundation Symposium on Equity and Educational Testing and Assessment, Washington, DC.